DOCUMENT RESUME

ED 262 057 TM 850 532

AUTHOR Frary, Robert B.; Olson, George H.
TITLE Detection of Coaching and Answer Copying on Standardized Tests.
INSTITUTION Virginia Polytechnic Inst. and State Univ., Blacksburg.
PUB DATE Apr 85
NOTE 9p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 1-3, 1985).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Cheating; Cluster Analysis; Elementary Secondary Education; Examiners; Quality Control; Scores; *Standardized Tests; *Test Coaching; *Testing Problems; Test Items; Test Results; Vectors (Mathematics)
IDENTIFIERS Test Batteries

ABSTRACT

Information on statistical detection of two different kinds of unethical activity associated with standardized testing is presented. The first is advance presentation of specific test items or excessive coaching using items very similar to those on a test. Cluster analysis is recommended for detection of aberrant vectors of percentages correct on the items of a test across testing units, such as classrooms or schools. The second activity is answer copying among examinees, which may reflect inadequate test administration procedures or outright encouragement of copying. Such activity can be detected by determining for each pair of examinees in a testing room, a statistic reflecting the extent to which their observed answer correspondences depart from expectation. The distribution of these statistics is normal in the absence of extensive copying, but becomes skewed when this phenomenom occurs. Examples of the application of the recommended methodologies in large school districts on the elementary and secondary levels are given. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED262057

Detection of Coaching and Answer Copying on Standardized Tests

Robert B. Frary, Virginia Polytechnic Institute State University
Blacksburg, Virginia 24061

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Frary, Robert B

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☑ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

TM 850 532

Detection of Coaching and Answer Copying on Standardized Tests

Robert B. Frary, Virginia Polytechnic Institute State University
Blacksburg, Virginia 24061

George H. Olson, Dallas Independent School District
Dallas, Texas 75204

Abstract: This paper presents information on statistical detection of two different kinds of unethical activity associated with standardized testing. The first is advance presentation of specific test items or excessive coaching using items very similar to those on a test. Cluster analysis is recommended for detection of aberrant vectors of percentages correct on the items of a test across testing units, such as classrooms or schools. The second activity is answer copying among examinees, which may reflect inadequate test administration procedures or outright encouragement of copying. Such activity can be detected by determining for each pair of examinees in a testing room, a statistic reflecting the extent to which their observed answer correspondences depart from expectation. The distribution of these statistics is normal in the absence of extensive copying, but becomes skewed when this phenomenon occurs. Examples of the application of the recommended methodologies are given.

Paper presented at the annual
meeting of the National Council
on Measurement in Education
Chicago, Illinois
April, 1985

Detection of Coaching and Answer Copying on Standardized Tests

Robert B. Frary, Virginia Polytechnic Institute and State University

George H. Olson, Dallas Independent School District

This paper presents information on statistical detection of two rather different kinds of unethical activity associated with standardized testing. These activities have a common goal, however. That is to obtain score averages meaningfully higher than would otherwise occur.

Unethical Coaching

The activity we will discuss first is advance presentation of specific test questions or excessive coaching using items very similar to those on a test. The statistical procedures we will recommend, while powerful, are not infallible. Therefore, we want to make clear at the outset that these procedures should not be applied in a vacuum; instead they can yield strong corroborative evidence to bolster that arising from the usual kinds of administrative surveillance required for any well-run testing system. In addition, they may suggest the need for closer surveillance of specific testing units such as classrooms, buildings or subdistricts.

The basic statistical procedure for detection of coaching is cluster analysis. This procedure groups together entities with similar values across a number of variables. For example, if the entities are students and the variables are scores on an achievement battery, one cluster might be students with high scores on all tests of the battery; another cluster might be students with high scores on verbal tests and moderate scores on quantitive tests.

To see how this clustering might detect unethical coaching, suppose that a school district has 50 elementary schools. These are the entities to be clustered. The variables will be the percentages of correct responses on the items of an achievement test within a battery. Depending on the clustering method used and the ability levels of the schools, only two or three clusters should be found. These should represent relatively high and low (or high, medium and low) achievement across all items of the test. In other words, the strongest schools should be best on all (or almost all) the items, and the weakest worst on all the

items. Suppose, say in a weaker school, unethical coaching is practiced. Then on some, but not all of the items, the percentages correct will approach those of the better schools. Then the offending school's vector of percentages correct will probably not resemble that of any other school; it will constitute a cluster of its own. Moreover, the content of the actual coaching is revealed for comparison with other evidence.[1]

It is important to note that this methodology depends on the difficulty of coaching students effectively over an entire test, much less a battery. Indeed, if it were possible to coach students effectively over an entire test, this outcome might well be considered tantamount to teaching. Thus, the successful result of coaching is aberrant percentages correct: for example, a school might do about as well as expected on half the items of a test and much better on the other half. This outcome would be especially revealing if the better half consisted entirely of items of the same type or format.

Cluster analysis is explained conceptually very well in an older review article by Baker (1972). Since that time the procedure has become available with several variations in statistical packages such as the Statistical Analysis System (SAS Institute, 1982) and BMDP Statistical Software (University of California Press, 1981). For data such as that just discussed almost any of the hierarchical clustering procedures is satisfactory, since only radical departures from expectation are of concern. The Johnson MAX procedure (see Baker, 1972) seems to work well with squared euclidean distance between vectors defining their degree of similarity. It is not necessary to standardize the percentages correct because they all measure on the same scale.

Sometimes a district is so large that the task of inspection for aberrant schools across the items of many tests becomes burdensome. In this case, a screening approach is desirable. Then the vector of interest is mean scores for a school or other testing unit (possibly classrooms) across all tests of a battery.[2] Just as coaching does not uniformly yield success across all items of a test, so it does not on all tests of a battery. Not only is time not likely to permit such coverage, but some areas are more sensitive to coaching than others. Schools with aberrant vectors of means can then be grouped with small numbers of similar but nonaberrant schools for comparison on an items basis on tests previously identified as suspect.

---

[1] In some cases this analysis will reveal curriculum or teaching problems unrelated to coaching. For example, performance much lower than expected on isolated items may show that specific material was not covered in a testing unit.

[2] In this case, the means should be standardized across schools before they are entered into the cluster analysis.

Answer Copying

At first consideration, answer copying might seem a problem in ethics of individual students but not one that threatens the validity of a testing program. This conclusion is correct if the answer copying is an isolated phenomenon. However, few administrators of testing programs have given proper consideration to the possible effect of large-scale copying. In fact, arranging for, encouraging or simply permitting a large proportion of students to copy answers or consult each other on test questions is an extremely effective way of increasing scores. Moreover, it is not detectable by the basic cluster-analytic procedure described above, because it affects no particular subset of items.[3]

This practice is insidious because it is difficult to confirm after-the-fact by interviewing examinees or observers. With younger examinees the test administrator need only create the atmosphere for cooperative group work and fail to stop the resulting collaboration. For older students, a slightly irreverant or lackadaisical attitude on the part of the test administrator will signal that cheating will not be punished. Then those who typically do so, given the opportunity, will. When others observe this activity, many may also try to enhance their scores. Even if extensive answer copying occurs in the absence of intentional efforts by school personnel to raise scores, its presence nevertheless reflects unsatisfactory testing conditions and inflated scores for some examinees.

Detection of answer copying is possible in a manner far more conclusive than discussed earlier with respect to aberrant patterns of item response. Responses of all pairs of examinees within a school or within any group tested together can be analyzed for unusual degrees of correspondence as outlined by Frary, Tideman and Watts (1977). For each pair of examinees a statistic is computed reflecting the extent to which their observed degree of correspondence deviates for expectation. Both right and wrong answers contribute to this statistic, which controls both for the score level of each examinee and the relative popularity of their right and wrong choices. When answer copying is extensive, the distribution of this statistic (called g2) becomes skewed due to excessive numbers of high values.

The g2 has been used extensively at Virginia Tech. One case started with an instructor's wish to determine the extent to which various security measures were preventing cheating on examinations in a multi-section course. The examination chosen for analysis had 30 4-choice items responded to by 944 examinees

---

[3]If practiced on a single or a small number of tests in a battery, the resulting means would appear aberrant if test means were cluster-analyzed.

in 11 rooms, all of whom took the test simultaneously. There were two forms of the test, designated A and B, the second a scrambled version of the first, and these were passed out alternately in each room.

Two distributions of g2's were produced for mutually exclusive groups:

Group 1: Pairs of examinees who were in the same room and took the same form of the test. There were 29,617 such pairs, which yielded 59,234 g2's, since the computation of a g2 changes when the potential copier and the person copied from are interchanged.

Group 2: Pairs of examinees who were in different rooms and took different forms of the test. A total of 192,816 such pairs were available, which yielded 385,632 g2's.

If it can be assumed that there was no crossform-crossroom cheating, the Group 2 distribution of g2's may be taken as a norm with which to compare other distributions arising from other subsamples of examinees on the same test. The highest g2 for the different form and room group was 4.60. The distribution of the same room and form group had 95 g2's above 4.60. This outcome suggested that there was extensive copying among about 5% of examinees. Further examination of the same room and form g2's revealed that, in six of the 11 testing rooms, no g2's above 4.60 were observed. These rooms were all relatively small, which suggests that the prevalence of cheating was related to room size.

Numerous other applications of the g2 statistic have been made. For example, three instructors each teaching a different section of the same course made different comments about cheating when passing out their midterm examination. One metioned nothing about cheating. The second stated that students should avoid looking at each other's responses. The third said that all pairs of papers would be checked for answer similarity. There was no evidence of cheating in this case and very little in the second. However, for the first case, cheating was rampant by any standard.

Standardized Testing Applications

The two procedures just described have been applied to standardized testing in two instances, both in large school districts. In the first case one elementary school was identified where unethical coaching had almost surely occurred. However, in secondary schools, where students were tested in large groups, answer copying was rampant in a number of schools, mostly senior high schools. In most of these cases the copying appeared to be the result of careless testing procedures rather than an overt

effort to raise scores. Nevertheless, the scores in question were undoubtedly higher as a result of the copying.

The second application was in the Dallas Independent School District and covered tests administered during the 1983-84 school year. No widespread evidence of answer copying or coaching was found. However, there were isolated instances of probable irregularities. One elementary school had excessive numbers of high g2 statistics on two of the tests administered. These same two tests yielded aberrant vectors of percentages correct across items. This outcome illustrates the fact that the two methods of detecting irregularities are not totally independent. Coaching that is effective for some but not all students could mimic the result of answer copying. However, as noted earlier, the reverse relationship is less plausible; answer dissemination through copying tends to affect all test items relatively the same. In this particular case, inappropriate provision of information to examinees probably occurred. Twelve of 60 items on a reading test had excessive percentages of correct answers. In contrast, the other test, in mathematics, had five items with substantially lower percentages answered correctly than at similar schools. No curricular deficiency or inadequate teaching could explain these deficits. On one item, only 19% answered correctly compared with about 70% at similar schools. Thus some sort of intervention seems to have occurred, perhaps even the provision of wrong answers or faulty methods which conflicted with earlier instruction.

At the time, schools in Dallas were competing for merit awards applicable to the salaries of all teachers in schools attaining higher levels of score growth. The school in question did not achieve the required level. However, if it had, the above information might have had serious impact.

The other irregularities noted in Dallas were isolated instances of extensive answer copying. These occurred in middle schools and in all cases involved some tests and not others in a battery. Thus, these cases probably resulted from careless test administration procedures rather than organized efforts to promote copying. In a large number of schools there was evidence of small numbers of examinees copying (exceedingly high g2 statistics for isolated pairs of examinees), which also suggested the need for better administration procedures.

Discussion

Although the methodology just described is straightforward and effective and has been available for a number of years, it has had little use. Factors influencing this state of affairs are numerous. Many school systems lack the statistical expertise and computer resources needed for this kind of quality control check. Among these are a number that lack the financial resources to

hire consultants to do the job on a regular basis. However, we believe the most prevalent reason for lack of use is the belief that "We don't have that problem and if we do it is minimal."

Even if this evaluation is correct, there is a great deal to recommend systematic application of the methodology we have described:

1. It is preventive. Those involved are unlikely even to consider unethical actions if they know there is a regular check for their occurrence.

2. It demonstrates the administration's commitment to quality. Many undesirable situations arise in part because no one appears concerned.

3. Irregularities, should they occur, are far more convincing as evidence when they follow years of data without such aberrations.

Independent of these considerations, some school systems avoid quality checks on their testing simply because they fear revelation of deficiencies. While no one would recommend this justification publicly, we believe it is a nontrivial influence against use of statistical methodology in this area.

In conclusion we strongly recommend regular and thorough statistical analyses of standardized test results to detect coaching and answer copying. With more widespread application of the statistical methodologies we have reviewed, improvements in their sensitivity and efficiency are likely outcomes. More important, we believe the effect on standardized testing would be substantially beneficial, both in quality of the resulting scores and in the increased confidence of those who interpret and use the scores.

References

Baker, F. B. (1972). Numerical taxonomy for educational researchers. Review of Educational Research, 42, 345-358.

Frary, R. B., Tideman, T. N., & Watts, T. M. (1977). Indices of cheating on multiple-choice tests. Journal of Educational Statistics, 2, 235-256.

SAS Institute, Inc. (1982). SAS user's guide: Statistics (1982 edition). Cary, NC: Author.

University of California Press. (1981). BMDP statistical software 1981. Berkley: Author.